Phi-Actor-Critic: Steering General-Sum Games to Pareto-Efficient Correlated Equilibria

Wongyu Lee[1] , **Francesco Lelli**[2] , **Omran Ayoub**[3] and **Massimo Tornatore**[1]
[1]Politecnico di Milano
[2]Tilburg University
[3]University of Applied Sciences and Arts of Southern Switzerland
wongyu.lee@polimi.it, f.lelli@tilburguniversity.edu, omran.ayoub@supsi.ch, massimo.tornatore@polimi.it

Abstract

Real-world multi-agent systems, from traffic coordination to resource allocation, are often modeled as general-sum games where individual incentives conflict with collective welfare. In these settings, the central challenge is not merely finding an equilibrium, but selecting socially desirable outcomes among many suboptimal Nash equilibria. Standard deep multi-agent reinforcement learning (MARL) methods struggle with this problem, as value-decomposition approaches are constrained by monotonicity assumptions and policy-gradient methods often converge to stable but socially inefficient equilibria. To address this limitation, we propose Φ-Actor-Critic (Φ-AC), a framework that leverages swap regret minimization to steer learning toward high-welfare correlated equilibria (CE). To make counterfactual regret estimation tractable in deep MARL, Φ-AC employs a centralized attention critic that predicts vector-valued regrets in a single forward pass, avoiding computationally expensive counterfactual simulations. We further introduce a Lagrangian-based equilibrium selection mechanism that optimizes social welfare while enforcing stability through regret constraints. Experiments on matrix games, Multi-Agent Particle Environments (MPE), and the Melting Pot Harvest scenario demonstrate that Φ-AC learns efficient and stable coordination strategies across diverse mixed-motive settings while maintaining high collective return and competitive fairness.

1 Introduction

Achieving socially efficient coordination, where agents collectively obtain higher welfare, remains challenging in multi-agent systems with partially conflicting incentives. Even when higher-welfare collective strategies exist, independent learning dynamics in mixed-motive environments often converge to individually rational but socially inefficient behaviors, such as risk-averse coordination or persistent asymmetric roles [Shoham and Leyton-Brown, 2008]. Existing deep multi-agent reinforcement learning (MARL) methods often struggle to avoid such outcomes. Value-decomposition approaches such as QMIX [Rashid *et al.*, 2020] rely on restrictive monotonicity assumptions, while policy-gradient methods (PGMs) such as MAPPO [Yu *et al.*, 2022] lack mechanisms that explicitly select socially desirable equilibria. As a result, these methods may converge to stable but low-welfare Nash equilibria (NE).

To address this limitation, we shift the focus from standard reward maximization to regret-based equilibrium selection rooted in algorithmic game theory (AGT) [Hart and Mas-Colell, 2000; Blum and Mansour, 2007; Cesa-Bianchi and Lugosi, 2006]. Rather than targeting only independent NE, regret-based learning enables correlated solution concepts that support richer coordination in general-sum games. In particular, while external regret minimization is associated with broad coarse correlated equilibria (CCE), swap-regret-based learning refines the target to the tighter class of correlated equilibria (CE) by evaluating whether an agent would benefit from conditionally replacing one action with another [Hart and Mas-Colell, 2000; Blum and Mansour, 2007].

Based on this insight, we propose Φ-Actor-Critic (Φ-AC), a regret-aware framework designed to operationalize swap regret for equilibrium selection in deep MARL. We address two practical challenges in applying swap regret to deep MARL: scalable regret estimation and equilibrium selection. First, since calculating swap regret typically requires computationally expensive counterfactual evaluation, we introduce a centralized attention critic that efficiently predicts vector-valued regrets in a single forward pass ($O(1)$). Second, since low regret only encourages behavior consistent with approximate CE conditions, it does not specify which equilibrium within the CE set should be selected. We therefore introduce the Regret-Balancing Social Welfare Objective (RB-SWO), which biases learning toward Pareto-efficient CE by optimizing social welfare under regret constraints. By formulating the learning problem as a constrained Markov game optimized via a Lagrangian objective, Φ-AC encourages socially efficient behavior while maintaining low swap regret.

Our contributions are threefold:

1. **Efficient Regret Estimation.** We introduce a regret-conditioned critic that predicts swap regret in a single forward pass, avoiding expensive counterfactual simulations in high-dimensional environments.
2. **Principled Equilibrium Selection.** We introduce a

Lagrangian-based selection mechanism that steers learning toward efficient and fair equilibria under regret constraints.

3. **Scalable Coordination in Social Dilemmas.** We demonstrate that Φ-AC learns high-welfare coordination strategies in matrix games and scales to complex Sequential Social Dilemmas (SSDs), outperforming strong MARL baselines in both sustainability and fairness.

2 Related Work

Scalable Deep MARL. Centralized-training decentralized-execution (CTDE) has become a standard paradigm for scalable MARL, allowing methods such as MADDPG [Lowe *et al.*, 2017] and COMA [Foerster *et al.*, 2018] to use centralized critics during training while preserving decentralized execution. Within this paradigm, COMA improves credit assignment through counterfactual baselines, while value-decomposition methods such as QMIX [Rashid *et al.*, 2020] further improve scalability by factorizing the joint action-value function through a monotonic mixing assumption. Recent PGM such as MAPPO [Yu *et al.*, 2022] also demonstrate strong empirical performance across cooperative MARL benchmarks. Despite these advances, these methods remain primarily optimized for expected return. QMIX is further constrained by its monotonic mixing assumption, which can be restrictive in mixed-motive games, while PGMs do not explicitly impose regret-based equilibrium selection. Thus, they may achieve stable learning or high returns without controlling whether the selected outcome is socially desirable.

Regret-Based Learning. Regret-minimization methods provide a principled route to equilibrium computation. Counterfactual Regret Minimization (CFR) and its neural variants, including Deep CFR and DREAM, minimize counterfactual regret and provide convergence guarantees primarily in two-player zero-sum imperfect-information games [Zinkevich *et al.*, 2007; Brown *et al.*, 2019; Steinberger *et al.*, 2020]. Similarly, NeuRD connects policy-gradient learning to Hedge [Freund and Schapire, 1997] and replicator dynamics in settings where the equilibrium objective is commonly formulated as approximate Nash convergence [Hennes *et al.*, 2020]. In contrast, Φ-AC targets general-sum MARL, where swap-regret-inspired learning naturally relates to correlated-equilibrium conditions and where the main challenge is selecting efficient CE rather than computing a zero-sum Nash solution.

Positioning of Φ-AC. Φ-AC bridges these lines of work by combining scalable centralized-critic MARL with swap-regret-based equilibrium selection. Rather than only improving return maximization or Nash convergence, Φ-AC uses learned swap-regret signals to steer policies toward Pareto-efficient CE while remaining applicable to high-dimensional mixed-motive environments.

3 Preliminaries

We formulate the environment as a general-sum Markov game and study equilibrium selection through the lens of CE.

3.1 General-Sum Markov Games

We model the multi-agent environment as a General-Sum Markov Game (MG), formally defined by the tuple $\mathcal{G} = \langle \mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \mathcal{P}, \{r_i\}_{i\in\mathcal{N}}, \gamma, \Omega, \mathcal{O}\rangle$. Here, $\mathcal{N} = \{1, \dots, N\}$ is the set of agents, $\mathcal{S}$ is the global state space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ denotes the state transition function, $r_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function for agent i, and $\gamma \in [0, 1)$ is the discount factor. We assume a partially observable setting where each agent i receives a local observation $o_i \in \Omega_i$ through the observation function $\mathcal{O} : \mathcal{S} \times \mathcal{N} \to \Omega_i$. The joint action is denoted by $\mathbf{a} = (a_i, \mathbf{a}_{-i}) \in \mathcal{A} \equiv \prod_i \mathcal{A}_i$.

3.2 Correlated Equilibrium and OSDP

While NE assumes independent policies, CE [Aumann, 1974] allows coordination via a correlation device, often enabling higher-welfare outcomes. In Markov games, however, equilibrium conditions involve deviations over sequential decision processes rather than a single normal-form interaction. The One-Shot Deviation Principle (OSDP) states that, under standard dynamic-game conditions, it is sufficient to check whether any agent can profit by deviating at a single decision point while following the original policy thereafter [Hendon *et al.*, 1996]. We use this principle as a local approximation of sequential deviation incentives. In our setting, this amounts to measuring whether an agent can improve its return by unilaterally changing its current action, using the stationary Q-function $Q_i^{\boldsymbol{\pi}}(s, \mathbf{a})$ as a local estimate of deviation incentives. We utilize the framework of Φ-Equilibria [Greenwald and Jafari, 2003]. Let Φ_i be a set of deviation functions $\phi : \mathcal{A}_i \to \mathcal{A}_i$ for agent i, where each deviation maps an action a_i to an alternative action $\phi(a_i)$.

Definition 1 (Φ-Equilibrium [Greenwald and Jafari, 2003])**.** A joint policy $\boldsymbol{\pi}$ constitutes a Φ-Equilibrium at state s if no agent i can improve their utility by applying any deviation $\phi \in \Phi_i$:

$$\sum_{\mathbf{a}\in\mathcal{A}} \boldsymbol{\pi}(\mathbf{a}|s) Q_i^{\boldsymbol{\pi}}(s, \mathbf{a}) \geq \sum_{\mathbf{a}\in\mathcal{A}} (\phi \circ_i \boldsymbol{\pi})(\mathbf{a}|s) Q_i^{\boldsymbol{\pi}}(s, \mathbf{a}), \quad \forall i \in \mathcal{N}, \forall \phi \in \Phi_i. \quad (1)$$

Here, $(\phi \circ_i \boldsymbol{\pi})$ denotes the joint distribution where agent i follows the deviation ϕ while others follow $\boldsymbol{\pi}$. When Φ_i includes all possible swap deviations, Eq. (1) characterizes CE.

3.3 Convergence via No-Swap-Regret Dynamics

A fundamental result in AGT links regret minimization to equilibrium convergence [Blum and Mansour, 2007; Cesa-Bianchi and Lugosi, 2006]. We define the instantaneous swap regret vector following standard swap-regret formulations [Hart and Mas-Colell, 2000; Blum and Mansour, 2007]. $\boldsymbol{\mathcal{R}}_i^{\Phi}(s, \mathbf{a}) \in \mathbb{R}^{|\mathcal{A}_i|}$ where each element corresponds to the counterfactual gain of swapping the chosen action a_i to a specific alternative $a_i' \in \mathcal{A}_i$:

$$\boldsymbol{\mathcal{R}}_i^{\Phi}(s, \mathbf{a})[a'] = \left[Q_i^{\boldsymbol{\pi}}(s, (a', \mathbf{a}_{-i})) - Q_i^{\boldsymbol{\pi}}(s, \mathbf{a})\right]^+. \quad (2)$$

where $[\cdot]^+$ denotes the ReLU operator. This vector formulation allows the critic to estimate regret for all possible deviations simultaneously, which is crucial for scalable optimization.

Proposition 1 (No-Φ-Regret to Approximate Φ-Equilibrium [Hart and Mas-Colell, 2000; Greenwald and Jafari, 2003; Blum and Mansour, 2007])**.** *Under standard finite repeated-game dynamics, let $x_t \in \Delta(\mathcal{A})$ be the joint action distribution induced at round t, and let $u_i(x_t)$ denote agent i's expected utility under x_t. Define $R_T^{\Phi,i} = \max_{\phi\in\Phi_i} \sum_{t=1}^{T} [u_i(\phi \circ_i x_t) - u_i(x_t)]$. If every agent satisfies $R_T^{\Phi,i}/T \leq \epsilon$, then the empirical distribution $\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t$ is an ϵ-Φ-equilibrium. When Φ_i contains all swap deviations, $\bar{x}_T$ is an ϵ-CE.*

The Selection Gap. While Proposition 1 establishes that minimizing swap regret leads to approximate CE behavior, it does not determine which equilibrium is selected. In general-sum games, some equilibria may still be socially inefficient or unfair, necessitating an explicit selection mechanism. In Section 4, we address this challenge via a constrained Markov game formulation using a Lagrangian objective that favors socially efficient equilibria.

4 Methodology: Φ-Actor-Critic

We propose Φ-AC, a deep MARL framework designed to encourage efficient coordination through swap-regret-based learning. The framework consists of three components: scalable regret estimation (Sec. 4.1), constrained equilibrium selection (Sec. 4.2), and exploration dynamics (Sec. 4.3). The overall centralized training optimization orchestrating these components is summarized in Algorithm 1.

4.1 Scalability: The Regret-Conditioned Critic

From Simulation to Prediction. The theoretical convergence to CE relies on calculating *Instantaneous Swap Regret* (Eq. 2). In tabular settings with a simulator, one can simply reset the environment to query "what if" for every unchosen action ($O(N|\mathcal{A}|)$ complexity). However, in real-time deployment where resetting is impossible, explicitly computing these counterfactual rewards becomes prohibitively expensive.

Φ-AC solves this bottleneck by training a centralized attention mechanism inspired by MAAC [Iqbal and Sha, 2019] to explicitly *predict* these counterfactual values. This provides dense regret estimates through a single forward pass, improving scalability in high-dimensional environments. As depicted in Figure 1, we design a dual-head critic to jointly model value estimation and swap-regret prediction. To account for non-stationary learning dynamics induced by concurrently adapting agents, we condition the critic on running cumulative regret $\boldsymbol{\mathcal{R}}^{\text{cum}}$ via a Feature-wise Linear Modulation (FiLM) layer, which dynamically scales and shifts critic features [Perez *et al.*, 2018].

The critic consists of two heads:

- **Q-Head:** Estimates action-values $Q_i^{\psi}(s, \mathbf{a}, \boldsymbol{\mathcal{R}}^{\text{cum}})$.
- **Regret-Head:** Predicts the regret vector $\hat{\boldsymbol{\mathcal{R}}}_i^{\psi} \in \mathbb{R}^{|\mathcal{A}_i|}$, where each element approximates the counterfactual gain of switching from a_i to a_i'.

To stabilize regret estimation, we compute the regression target $\boldsymbol{\mathcal{R}}_i^{\text{target}}$ using a Target Network ψ' with parameters lagging ψ:

$$\boldsymbol{\mathcal{R}}_i^{\text{target}}(a') = \left[Q_i^{\psi'}(s, (a', \mathbf{a}_{-i})) - Q_i^{\psi'}(s, \mathbf{a})\right]^+. \quad (3)$$

The critic minimizes the joint loss, including $\mathcal{L}_R = \frac{1}{N}\sum_i \|\boldsymbol{\mathcal{R}}_i^{\text{target}} - \hat{\boldsymbol{\mathcal{R}}}_i^{\psi}\|_2^2$.

Algorithm 1 Centralized Training Optimization in Φ-AC

Input: Minibatch $\mathcal{B} = (\mathbf{o}, \mathbf{a}, \mathbf{r}, \mathbf{o}')$ from $\mathcal{D}$, Current Cumulative Regret $\boldsymbol{\mathcal{R}}^{\text{cum}}$.
Hyperparameters: δ_{regret} (Tolerance), δ (EMA Decay rate), η_α (Dual LR).
// 1. Regret-Conditioned Critic Update (Sec. 4.1)
Calculate TD targets y_i and regret targets $\boldsymbol{\mathcal{R}}_i^{\text{target}}$ using target networks ψ', θ'.
Get outputs from dual-head Critic conditioned on $\boldsymbol{\mathcal{R}}^{\text{cum}}$ via FiLM:
  $\{Q_i^{\psi}, \hat{\boldsymbol{\mathcal{R}}}_i^{\psi}\} \leftarrow \text{Critic}(\mathbf{o}, \mathbf{a}, \boldsymbol{\mathcal{R}}^{\text{cum}})$
Update Critic ψ by minimizing the joint loss:
  $\mathcal{L}_{\text{Critic}} = \frac{1}{N}\sum_i \left(\mathcal{L}_Q(Q_i^{\psi}, y_i) + \mathcal{L}_R(\hat{\boldsymbol{\mathcal{R}}}_i^{\psi}, \boldsymbol{\mathcal{R}}_i^{\text{target}})\right)$
// 2. Actor & Dual Update via RB-SWO (Sec. 4.2)
  Sample differentiable actions $\tilde{\mathbf{a}} \sim \boldsymbol{\pi}_\theta(\mathbf{o})$ via Gumbel-Softmax using regret-biased logits (Eq. (10))
Get Critic predictions $\{Q_i^{\psi}(\tilde{\mathbf{a}}), \hat{\boldsymbol{\mathcal{R}}}_i^{\psi}(\tilde{\mathbf{a}})\}$ evaluated at the sampled actions.
Calculate Regret Magnitudes $\mathcal{M}_i \leftarrow \|[\hat{\boldsymbol{\mathcal{R}}}_i^{\psi}(\tilde{\mathbf{a}})]^+\|_2$.
[Primal] Update Actor θ by minimizing:
  $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{RB-SWO}}(\theta, \boldsymbol{\alpha})$ (Eq. (8))
[Dual] Update multipliers $\alpha_i^{\text{fair}}, \alpha_i^{\text{ent}}$ via gradient ascent:
  $\alpha_i^{\text{fair}} \leftarrow \left[\alpha_i^{\text{fair}} + \eta_\alpha(\mathcal{M}_i - \delta_{\text{regret}})\right]^+$
  $\alpha_i^{\text{ent}} \leftarrow \left[\alpha_i^{\text{ent}} + \eta_\alpha(\mathcal{H}_i^{\text{target}} - \mathcal{H}(\pi_{\theta_i}))\right]^+$
// 3. Cumulative Regret & Target Networks Update (Sec. 4.3)
Update Cumulative Regret using Exponential Moving Average (EMA):
  $\boldsymbol{\mathcal{R}}_i^{\text{cum}} \leftarrow \delta\boldsymbol{\mathcal{R}}_i^{\text{cum}} + (1-\delta)\hat{\boldsymbol{\mathcal{R}}}_i^{\psi}(\mathbf{a})$
Soft update target networks ψ' and θ'.

4.2 Selection: Lagrangian Dual Optimization

Minimizing swap regret encourages stability, but it does not determine which equilibrium is selected when multiple low-regret equilibria exist. To address this selection gap, we introduce the Regret-Balancing Social Welfare Objective (RB-SWO), which maximizes collective welfare while constraining each agent's regret magnitude. This formulation favors high-welfare coordination without allowing the policy to move far from the low-regret region.

Formulation as Constrained Optimization. We formalize RB-SWO as a constrained optimization problem over expected discounted returns. We explicitly define the *Regret Magnitude* $\mathcal{M}_i(\boldsymbol{\pi})$ as the expected L_2-norm of the positive swap regret vector:

$$\mathcal{M}_i(\boldsymbol{\pi}) = \mathbb{E}\left[\left\|\left[\hat{\boldsymbol{\mathcal{R}}}_i^{\psi}(s, \mathbf{a})\right]^+\right\|_2\right] \quad (4)$$

We also define the policy entropy as $\mathcal{H}(\pi_i) = \mathbb{E}_{a_i\sim\pi_i}[-\log \pi_i(a_i|o_i)]$, with $\mathcal{H}_i^{\text{target}}(t)$ denoting a time-varying minimum entropy target.

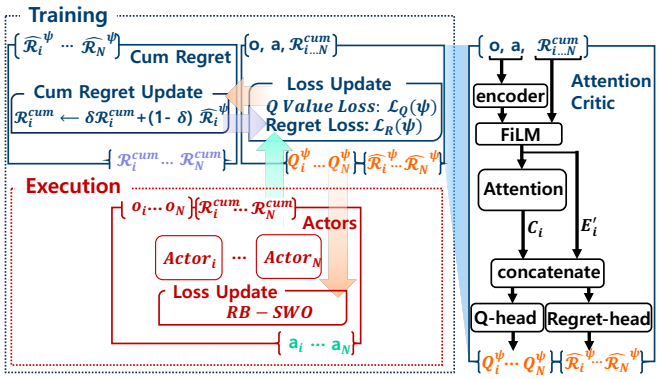


Figure 1: The architecture of Φ-AC. (Left) Decentralized actors select actions using local observations and cumulative regret bias. (Right) The centralized attention critic processes global states and cumulative regret to output both Q-values and explicit regret predictions. (Center) The Regret Coordinator modulates the critic via FiLM based on the cumulative regret feedback loop.

The optimization problem is:

$$\max_{\boldsymbol{\pi}} \quad J(\boldsymbol{\pi}) = \mathbb{E}_{s\sim d^{\boldsymbol{\pi}},\mathbf{a}\sim\boldsymbol{\pi}}\left[\sum_{i\in\mathcal{N}} Q_i^{\boldsymbol{\pi}}(s,\mathbf{a})\right] \tag{5}$$

$$\text{s.t.} \quad \mathcal{M}_i(\boldsymbol{\pi}) \le \delta_{\text{regret}}, \quad \forall i \in \mathcal{N} \tag{6}$$

$$\mathcal{H}(\pi_i) \ge \mathcal{H}_i^{\text{target}}(t), \quad \forall i \in \mathcal{N}. \tag{7}$$

where δ_{regret} controls the allowed regret magnitude, and $\mathcal{H}_i^{\text{target}}(t)$ is annealed over training. We employ the Gumbel-Softmax reparameterization trick [Jang *et al.*, 2016] to enable differentiable optimization of the regret constraint.

Lagrangian Dual Optimization. We solve this via the primal-dual method. Introducing the dual variables $\boldsymbol{\alpha} = \{\alpha_i^{\text{fair}}, \alpha_i^{\text{ent}}\}_{i=1}^N$, where $\alpha_i^{\text{fair}} \ge 0$ controls regret-based regularization and $\alpha_i^{\text{ent}} \ge 0$ encourages exploration, we minimize the following objective for actor parameters θ:

$$\mathcal{L}_{\text{total}}(\theta,\boldsymbol{\alpha}) = \mathbb{E}\Bigg[\underbrace{-\sum_i Q_i^{\boldsymbol{\pi}}}_{\text{Max Welfare}} + \sum_i \alpha_i^{\text{fair}}(\mathcal{M}_i - \delta_{\text{regret}}) + \sum_i \alpha_i^{\text{ent}}(\mathcal{H}_i^{\text{target}} - \mathcal{H}(\pi_i))\Bigg]. \tag{8}$$

The optimization proceeds as an alternating game:

- **Primal Update (Actor):** Update θ to minimize $\mathcal{L}_{\text{total}}$.
- **Dual Update (Multipliers):** Update multipliers via gradient ascent with step size $\eta_\alpha > 0$:

$$\alpha_i^{\text{fair}} \leftarrow \left[\alpha_i^{\text{fair}} + \eta_\alpha(\mathcal{M}_i - \delta_{\text{regret}})\right]^+. \tag{9}$$

When regret increases, the corresponding dual variable also increases, placing greater emphasis on stability during optimization.

4.3 Dynamics: Entropy Annealing & Warm-up

A practical challenge in regret-based learning is that agents cannot regret actions they have never explored. To resolve this, we implement an explicit schedule:

Discovery (Warm-up). For the first T_{warm} steps (a predefined hyperparameter), we suppress the fairness penalty ($\alpha^{\text{fair}} \to 0$) and enforce a high entropy target. This forces agents to explore the state space purely via Maximum Entropy RL (MaxEnt RL) [Haarnoja *et al.*, 2018], accumulating the reward experiences necessary to define well-defined regret.

Convergence (Annealing). Post warm-up, we activate the regret constraints and linearly decay the entropy, encouraging diverse exploration before convergence to stable coordinated behaviors.

5 Theoretical Analysis

In this section, we provide a theoretical interpretation of Φ-AC by connecting its policy update to no-swap-regret learning and by explaining how the proposed Lagrangian objective biases equilibrium selection. Rather than relying on exact tabular regret computation, Φ-AC uses a learned critic to approximate regret signals in a differentiable deep MARL setting.

5.1 Approximating Regret Matching via Deep Learning

We next interpret the actor update in Φ-AC through the lens of regret matching.

Actor as Smooth Regret Matching. Classically, convergence to CE can be achieved via Regret Matching (RM) [Hart and Mas-Colell, 2000], where agents sample actions proportionally to their positive regret values, i.e., $\pi(a_i) \propto [\boldsymbol{\mathcal{R}}_i^{\Phi}(a_i)]^+$. However, the non-differentiability of the standard RM operator makes it unsuitable for gradient-based optimization in deep networks.

To bridge this gap, Φ-AC operationalizes RM via a Regret-Biased Softmax Policy. By injecting the predicted cumulative regret $\hat{\boldsymbol{\mathcal{R}}}^{\text{cum}}$ into the logits, the policy becomes:

$$\pi_\theta(a_i|o_i) = \frac{\exp\left((L_i(o_i,a_i) + \beta[\hat{\boldsymbol{\mathcal{R}}}_i^{\text{cum}}(a_i)]^+)/\tau\right)}{\sum_{a'}\exp\left((L_i(o_i,a') + \beta[\hat{\boldsymbol{\mathcal{R}}}_i^{\text{cum}}(a')]^+)/\tau\right)}. \tag{10}$$

Here, $L_i(o_i,a_i)$ denotes the raw logits output by the actor network prior to softmax normalization, τ is the temperature, and β controls the strength of the regret bias. Lower temperatures make the policy more sensitive to positive regret, while larger β places greater weight on actions with accumulated regret. Together, these terms yield a smooth and differentiable approximation to regret-driven action selection.

This formulation can be interpreted as a smooth approximation of RM. While standard RM selects actions in proportion to positive regret, our softmax-based formulation is conceptually related to exponential weights and Hedge algorithms [Freund and Schapire, 1997; Cesa-Bianchi and Lugosi, 2006], which are also known to induce no-regret behavior under standard online learning assumptions.

Approximate Regret under Function Approximation. In deep MARL, regret estimates are inherently approximate due to function approximation. Following standard analyses of perturbed no-regret dynamics [Cesa-Bianchi and Lugosi, 2006], we consider the setting where the centralized critic provides uniformly bounded regret estimation error: $\max_{s,\mathbf{a}} \|\mathcal{R}^{\text{true}} - \hat{\mathcal{R}}^{\psi}\|_\infty \leq \epsilon_c$. Under policy updates that are sufficiently smooth with respect to the regret estimates, the induced perturbation scales with ϵ_c, implying that the resulting average swap regret remains bounded up to an $\mathcal{O}(\epsilon_c)$ approximation term. Consequently, the learned dynamics may approach an approximate CE despite imperfect regret estimation.

5.2 Geometry of Equilibrium Selection

While regret minimization encourages behavior consistent with approximate CE conditions, it does not by itself distinguish high-welfare equilibria from inefficient low-regret outcomes. RB-SWO addresses this selection gap by adding welfare maximization under individual regret constraints, yielding a Lagrangian objective that favors high-welfare coordination.

Dual Forces in Optimization. We formulate equilibrium selection as maximizing social welfare ($W = \sum Q_i$) subject to stability constraints ($\mathcal{M}_i \leq \delta_{\text{regret}}$). The Lagrangian objective $\mathcal{L}(\theta, \alpha) = W - \sum \alpha_i(\mathcal{M}_i - \delta_{\text{regret}})$ induces two distinct gradient forces:

Efficiency Force ($\nabla_\theta W$): This term drives the joint policy along the "flat" directions of the CE polytope (where regret is zero) toward higher-welfare regions of the CE set.

Restoring Force ($-\alpha_i \nabla_\theta \mathcal{M}_i$): If an agent deviates from equilibrium (increasing $\mathcal{M}_i$), the dual variable α_i increases (Dual optimization). This increases the optimization pressure toward lower-regret regions.

Fairness via Individual Constraints. Individual regret constraints also play an important role in equilibrium selection. Geometrically, asymmetric equilibria (where one agent yields and another benefits) imply that while the sum of regrets might be low, the individual regret for the yielding agent is high. By enforcing $\mathcal{M}_i \leq \delta$ for each agent rather than only controlling the average regret, RB-SWO discourages solutions in which one agent absorbs most of the deviation incentive. This encourages more balanced outcomes and reduces winner-takes-all behavior.

6 Experimental Evaluation

While the previous sections introduced the regret-aware architecture and optimization framework of Φ-AC, this section empirically evaluates its ability to learn socially efficient coordination. We evaluate Φ-AC across a hierarchy of complexity: diagnostic matrix games, scalable MPE domains, and Melting Pot Harvest. These settings respectively test equilibrium selection in transparent games, scalability in higher-dimensional interactions, and sustainable coordination under shared resource depletion.

We compare Φ-AC against representative MARL baselines that capture different algorithmic design choices. MADDPG represents deterministic policy-gradient learning, MAPPO and COMA represent stochastic policy-gradient methods with centralized training, and QMIX represents value decomposition under monotonic factorization.

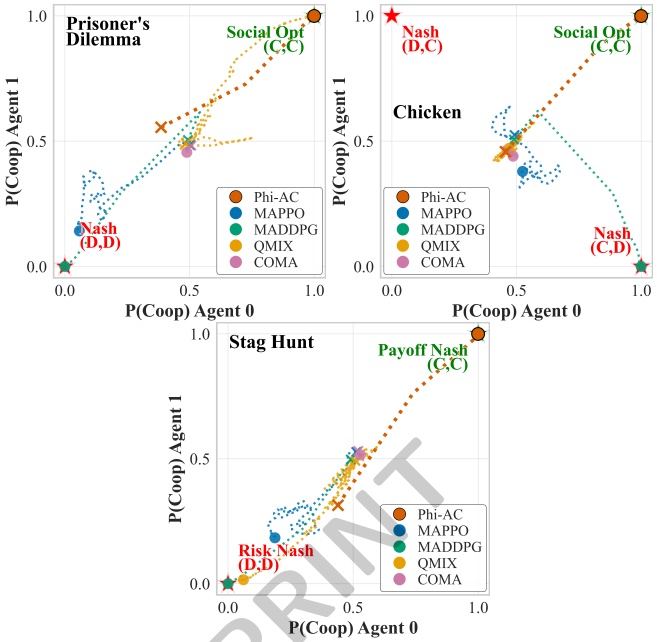


Figure 2: Policy Trajectories. (Top-L) Prisoner's Dilemma, (Top-R) Chicken, (Bot) Stag Hunt. Φ-AC (Red) consistently reaches the intended high-welfare equilibria.

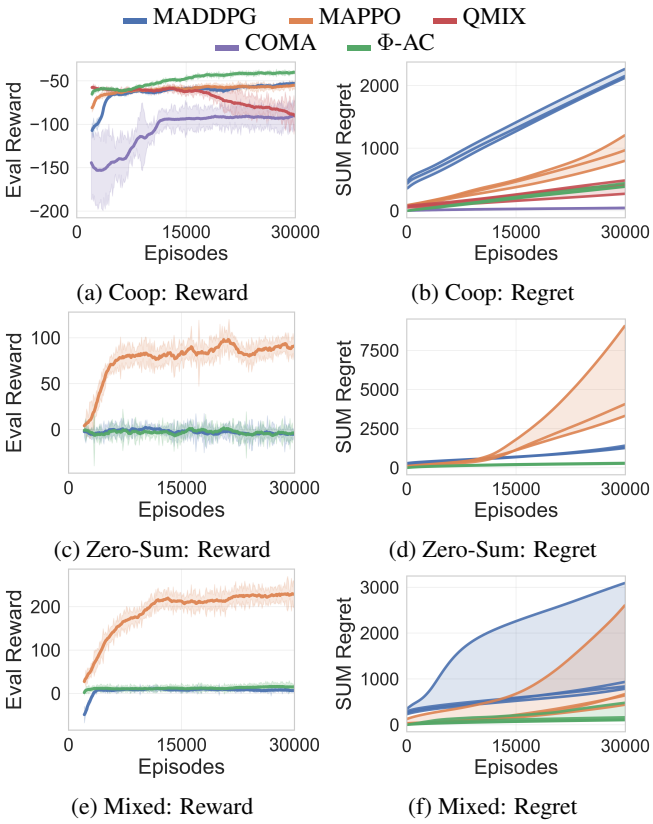
Experimental Setup. Training configurations were tailored to the complexity of each domain: Iterated Matrix Games (IMG) were trained for 1k episodes with a horizon of 25 steps, MPE environments [Lowe *et al.*, 2017] for 30k episodes (horizon 25), and Melting Pot (Harvest) [Leibo *et al.*, 2021] for 10k episodes (horizon 500). To ensure statistical robustness, all results for MPE and Melting Pot are reported as mean ± standard deviations over 5 independent runs.

6.1 Diagnostic Analysis: Iterated Matrix Games

We first employ Iterated Matrix Games to empirically verify the equilibrium convergence of each model. We consider a two-player setting where agents interact repeatedly, allowing them to adapt strategies based on history. We select three canonical games, each representing a distinct coordination failure mode. Fig. 2 visualizes the learning trajectories.

Prisoner's Dilemma. This game tests whether agents can overcome the temptation to defect and coordinate on mutual cooperation. As shown in Fig. 2 (Top-L), MADDPG and MAPPO tend to converge toward the inefficient NE (D,D), while QMIX and Φ-AC reach the social optimum (C,C). Unlike QMIX, which benefits from the monotonic payoff structure in this setting, Φ-AC reaches mutual cooperation through regret-aware coordination.

Chicken. This game tests coordination under two asymmetric equilibria, where either agent can benefit if the other yields. As shown in Fig. 2 (Top-R), baselines often drift toward one-



(a) Coop: Reward (b) Coop: Regret

(c) Zero-Sum: Reward (d) Zero-Sum: Regret

(e) Mixed: Reward (f) Mixed: Regret

Figure 3: Training Dynamics. Rows represent environments: Cooperative, Zero-Sum, and Mixed. Left panels show evaluation rewards, and right panels show cumulative regret.

sided or unstable outcomes, whereas Φ-AC promotes balanced coordination by reducing regret disparities between agents.

Stag Hunt. This game tests whether agents can coordinate on a high-payoff action that is beneficial only when both agents choose it, rather than falling back to the safer but lower-reward option. As shown in Fig. 2 (Bottom), baselines tend to favor the conservative risk-dominant outcome, while Φ-AC consistently reaches the payoff-dominant cooperative equilibrium.

6.2 Selection under Scale: MPE

We extend our evaluation to the MPE to verify whether the selection capability scales to high-dimensional state-action spaces via our centralized attention critic. We report Mean Reward (efficiency), Cumulative Regret (Convergence), and Regret Gap ($|\mathcal{R}_i^{\text{cum}} - \mathcal{R}_j^{\text{cum}}|$, fairness). Table 1 summarizes the quantitative results, and Figure 3 visualizes the training dynamics.

Cooperative: Avoiding Inefficient Equilibria. In `simple_spread`, agents must cover landmarks while avoiding collisions. COMA and QMIX achieve low regret but converge to lower-welfare solutions, suggesting that stability alone does not imply efficient equilibrium selection. In contrast, Φ-AC achieves the highest reward (-42.24). By balancing welfare maximization with regret constraints, it avoids premature stabilization and reaches a higher-welfare configuration.

Zero-Sum: Stability vs. Exploitation. High individual rewards in `simple_adversary` often signal exploitation rather than equilibrium. MAPPO achieves a high reward (91.17) but suffers from large regret gap (5926.82), indicating a non-stationary cycle where one agent continually exploits a weak opponent without convergence. Φ-AC converges to a reward near zero (-3.27) with significantly lower regret, suggesting a more stable low-regret regime in which unilateral exploitation incentives are reduced.

Mixed-Motive: Scalable Fairness. `simple_tag` involves conflicting goals within a general-sum framework. Baselines exhibit large regret gaps (> 2000), indicating highly imbalanced deviation incentives consistent with winner-takes-all behavior. Φ-AC reduces this gap by nearly two orders of magnitude (53.79). This suggests that the RB-SWO mechanism can scale to high-dimensional domains, enforcing equity by ensuring that the dissatisfaction (regret) is distributed equitably among agents.

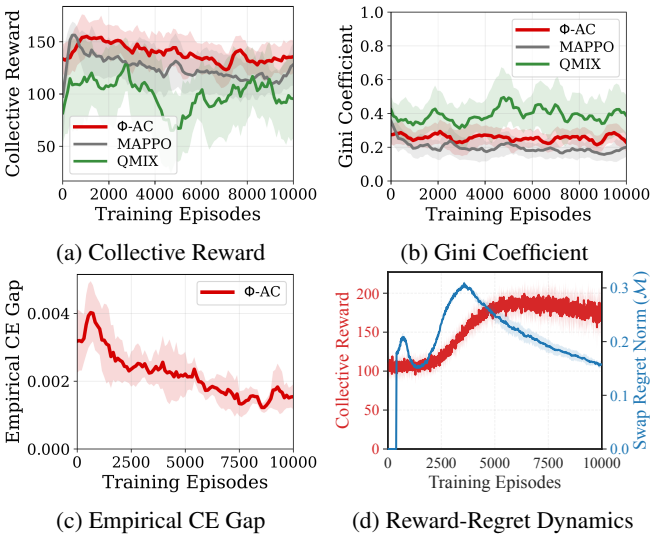
Model	Reward	Cum. Regret	Regret Gap
Cooperative (Spread)			
MADDPG	-53.55 ± 0.55	2177.97 ± 46.98	165.32 ± 61.98
MAPPO	-55.68 ± 3.11	988.00 ± 82.29	527.92 ± 272.01
COMA	-91.24 ± 13.08	$\mathbf{43.72 \pm 11.54}$	$\mathbf{39.93 \pm 25.37}$
QMIX	-88.49 ± 11.54	390.13 ± 41.66	248.44 ± 60.42
Φ-AC	$\mathbf{-42.24 \pm 0.21}$	460.46 ± 34.69	260.64 ± 119.98
Zero-Sum (Adversary)			
MADDPG	-4.91 ± 3.65	1316.86 ± 105.13	187.11 ± 122.92
MAPPO	91.17 ± 2.96	5470.50 ± 1845.02	5926.82 ± 6044.88
Φ-AC	$\mathbf{-3.27 \pm 1.63}$	$\mathbf{281.30 \pm 11.10}$	$\mathbf{43.32 \pm 18.28}$
Mixed-Motive (Tag)			
MADDPG	6.83 ± 0.93	1409.98 ± 50.54	2321.88 ± 140.22
MAPPO	227.69 ± 13.24	1084.68 ± 202.97	2192.30 ± 654.54
Φ-AC	$\mathbf{12.55 \pm 7.32}$	$\mathbf{481.56 \pm 20.51}$	$\mathbf{53.79 \pm 12.09}$

Table 1: Scalability results on MPE. Φ-AC achieves the best balance of efficiency and stability. Note: In zero-sum settings, a reward near 0 with low regret implies a stable NE, whereas high rewards indicate exploitation.

6.3 Ablation on Learning Objectives

We dissect the contribution of our proposed learning objectives in `simple_tag`. The results, summarized in Table 2, reveal that each component is critical for preventing unstable learning outcomes.

1. **w/o Regret Matching (RM) Bonus:** Removing the regret-bias term ($\beta = 0$) leads to unstable learning dynamics. The total regret increases substantially to **1933.04** with an extreme standard deviation ($\pm$ 2007.34). This indicates that without the RM bonus, the learning dynamics fail to converge, oscillating severely or diverging completely.
2. **w/o Fairness Objective (RB-SWO):** Removing the fairness constraint ($\alpha^{\text{fair}} \to 0$) degrades both stability ($481.56 \to 510.11$ in Regret) and efficiency. This suggests that the fairness objective acts as a dual-purpose reg-



(a) Collective Reward (b) Gini Coefficient

(c) Empirical CE Gap (d) Reward-Regret Dynamics

Figure 4: Dynamics in Harvest. (a,b) Φ-AC achieves high collective reward while maintaining competitive fairness. (c) The empirical CE-gap proxy remains low during evaluation. (d) Training reward-regret dynamics show decreasing regret alongside sustained collective reward.

ularizer: it not only prevents inefficient outcomes ("Coordination Collapse") but also accelerates convergence by constraining the search space to the stable region.

Model Variant	Final Reward	Total Regret
Full Φ-AC (Proposed)	**12.55 ± 7.32**	**481.56 ± 20.51**
w/o RM Bonus	12.53 ± 7.95	1933.04 ± 2007.34
w/o Fairness Objective	10.94 ± 7.19	510.11 ± 28.46

Table 2: Ablation results on `simple_tag`. Removing the RM Bonus leads to severe instability (large regret variance), while removing Fairness degrades both convergence and reward. Φ-AC achieves the most stable and efficient equilibrium.

6.4 Robustness in Dilemmas: Melting Pot

Finally, we evaluate Φ-AC on Harvest, an SSD that serves as a challenging benchmark for sustainable coordination under shared resource constraints. In this environment, apples regenerate based on the density of nearby unharvested apples, creating a "Tragedy of the Commons" where individually greedy behavior can undermine long-term collective productivity. We compare Φ-AC against MAPPO (centralized policy-gradient baseline) and QMIX (value decomposition baseline) using the Sustainability Index (SI), defined as $SI = (1 - \text{Gini}) \times \sum_i \text{Return}_i$, which captures the trade-off between productivity and fairness.

Escaping Resource Collapse (Fig. 4 (a,b)). As shown in Table 3, Φ-AC achieves the highest collective reward (134.33) while maintaining competitive fairness relative to MAPPO. In contrast, QMIX suffers from severe inequality and substantially lower collective productivity, resulting in the weakest Sustainability Index. Although MAPPO achieves the lowest Gini coefficient, its lower collective reward leads to a smaller overall sustainability score than Φ-AC. These results suggest that regret-based equilibrium selection helps maintain productive coordination without collapsing into resource-depleting behavior.

Reward–Regret Dynamics (Fig. 4 (c,d)). To further analyze the learned coordination dynamics, we additionally measure the empirical CE gap, defined as the maximum positive swap regret estimated by the critic during evaluation. As shown in Fig. 4(c), the empirical CE gap of Φ-AC remains consistently low throughout training, suggesting reduced unilateral deviation incentives among agents.

This interpretation is further supported by the reward-regret dynamics in Fig. 4(d). During early training, swap regret temporarily increases as agents explore alternative coordination strategies. As learning progresses, however, the regret signal decreases while collective reward remains high. This indicates that agents gradually settle into a stable high-welfare coordination pattern rather than maximizing reward through unstable exploitative behavior.

Model	Coll. Reward	Gini Coeff	Sust. Index
Φ-AC (Ours)	**134.33 ± 3.46**	0.26 ± 0.03	**99.41 ± 2.67**
MAPPO	116.55 ± 7.96	**0.18 ± 0.03**	95.48 ± 7.60
QMIX	94.49 ± 25.08	0.40 ± 0.05	59.27 ± 18.27

Table 3: Performance summary in harvest (Evaluation, Last 10%). Φ-AC achieves the highest Sustainability Index while maintaining strong collective productivity.

7 Conclusion

In this work, we introduced Φ-AC, a principled framework designed to integrate game-theoretic swap regret minimization with deep MARL. By shifting the paradigm from reward-only optimization to equilibrium selection, Φ-AC addresses the limitations of conventional MARL objectives in distinguishing efficient coordination from suboptimal stability.

Our empirical evaluation supports Φ-AC across a hierarchy of complexity. At the diagnostic level (Matrix Games), we visualized robust steering toward social optima. At the scalable level (MPE), we demonstrated effective fairness-aware coordination. Most significantly, in the complex SSD (Harvest), we provided empirical evidence of high-welfare coordination behavior consistent with correlated-equilibrium principles.

The observed decrease in smoothed swap regret alongside increasing collective reward suggests that Φ-AC promotes sustained coordination rather than short-term reward maximization. A current limitation is its reliance on discrete action spaces. Future work will extend regret estimation to continuous control, broadening game-theoretic equilibrium selection to physical decision-making domains.

References

[Aumann, 1974] Robert J Aumann. Subjectivity and correlation in randomized strategies. *Journal of mathematical Economics*, 1(1):67–96, 1974.

[Blum and Mansour, 2007] Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8(6), 2007.

[Brown *et al.*, 2019] Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. Deep counterfactual regret minimization. In *International conference on machine learning*, pages 793–802. PMLR, 2019.

[Cesa-Bianchi and Lugosi, 2006] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[Foerster *et al.*, 2018] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[Freund and Schapire, 1997] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[Greenwald and Jafari, 2003] Amy Greenwald and Amir Jafari. A general class of no-regret learning algorithms and game-theoretic equilibria. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, pages 2–12. Springer, 2003.

[Haarnoja *et al.*, 2018] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.

[Hart and Mas-Colell, 2000] Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

[Hendon *et al.*, 1996] Ebbe Hendon, Hans Jørgen Jacobsen, and Birgitte Sloth. The one-shot-deviation principle for sequential rationality. *Games and Economic Behavior*, 12(2):274–282, 1996.

[Hennes *et al.*, 2020] Daniel Hennes, Dustin Morrill, Shayegan Omidshafiei, Rémi Munos, Julien Perolat, Marc Lanctot, Audrunas Gruslys, Jean-Baptiste Lespiau, Paavo Parmas, Edgar Duéñez-Guzmán, et al. Neural replicator dynamics: Multiagent learning via hedging policy gradients. In *Proceedings of the 19th international conference on autonomous agents and multiagent systems*, pages 492–501, 2020.

[Iqbal and Sha, 2019] Shariq Iqbal and Fei Sha. Actor-attention-critic for multi-agent reinforcement learning. In *International conference on machine learning*, pages 2961–2970. PMLR, 2019.

[Jang *et al.*, 2016] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[Leibo *et al.*, 2021] Joel Z Leibo, Edgar A Dueñez-Guzman, Alexander Vezhnevets, John P Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charlie Beattie, Igor Mordatch, and Thore Graepel. Scalable evaluation of multi-agent reinforcement learning with melting pot. In *International conference on machine learning*, pages 6187–6199. PMLR, 2021.

[Lowe *et al.*, 2017] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

[Perez *et al.*, 2018] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[Rashid *et al.*, 2020] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020.

[Shoham and Leyton-Brown, 2008] Yoav Shoham and Kevin Leyton-Brown. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

[Steinberger *et al.*, 2020] Eric Steinberger, Adam Lerer, and Noam Brown. Dream: Deep regret minimization with advantage baselines and model-free learning. *arXiv preprint arXiv:2006.10410*, 2020.

[Yu *et al.*, 2022] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35:24611–24624, 2022.

[Zinkevich *et al.*, 2007] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. *Advances in neural information processing systems*, 20, 2007.